GeoBrain
A Web Service Based Geospatial Knowledge Building System

Liping Di
Laboratory for Advanced Information Technology and Standards (LAITS)
George Mason University
9801 Greenbelt Road, Suite 316-317
Lanham, MD 20706, USA
ldi@gmu.edu

NASA EOS Higher-Education Alliance: Mobilization of NASA EOS Data and Information Through Web Services and Knowledge Management Technologies for Higher-Education Teaching and Research

ESE Information Systems Vision for 2015

(From the capability vision workshop summary)

- **From ESE Strategy (Oct 2003)**
  - " Advanced information systems to enable the processing, communicating, and archiving of vast amounts of data generated by the envisioned networks of sensorcrafts, and to deliver on-demand and affordably Earth system information products to customers located anywhere and at anytime."
- **Working Vision**
  - Near-real-time, transparent, seamless, and automatic...
  - data fusion, data analysis, and knowledge discovery...
  - from petabytes of data acquired from multiple sources...
  - to enable and accelerate progress toward ESE goals for scientific research, applications, and education.

The Objectives of the Research

- To enable the students and faculty of higher-education institutes easily accessing, analyzing, and modeling with the huge volume of NASA EOSDIS data for teaching and research just like they possess such vast resources locally at their desktops.
  - Enable the education users to handle vast NASA EOS data and computing resources like their local ones.
  - Develop/enhance courses that fully utilize the environment for Earth System Science/Geospatial education
- To realize this goal, we will develop an open, standard-based interoperable web geospatial information system called *GeoBrain* and operate it on top of NASA EOSDIS on-line data resources
  - Develop geospatial web service and knowledge management technologies for NASA EOS data environment.
  - Implement them in an open, standard-based, distributed, interoperable web service system.
  - It is a geospatial modeling and knowledge building system

Process of Learning and Knowledge Discovery in Data-Intensive ESS

1. Find a real-world problem to solve
2. Develop/modify a hypothesis/model
3. Implement the model/develop analysis procedure at computer systems.
4. Determine the data requirements and search, find, and order the data from data providers.
5. Preprocess the data into the ready-to-analysis form
   - reprojection, reformating, subsetting, subsampling, geometric/radiometric correction, etc.
6. Execute the model/analysis procedure to obtain the results.
7. Analyze and validate the results
8. Repeat steps 2-7 until the problem is solved.

ESS Data Available at NASA

- The NASA Earth Observing System (EOS) collects more than 2Tb of remote sensing data/ day.
- Currently NASA Active Archive Data Centers (DAACs) have archived multiple peta bytes of data from EOS and pre-EOS era.
  - Significant part of the data archives have never been analyzed once.
- All of those data are free to all data users.

NASA ESS Data Environment

- The EOS data and information system (EOSDIS) is designed to manage, archive, analyze, and distribute the ESS data.
  - Originally designed for supporting NASA funded scientists.
  - Based on technologies of 20 years ago.
  - Mainly for supporting well-funded NASA ESS research projects
  - Not considering the small data users and educators.
- The standard data format in EOSDIS is HDF-EOS.
- EOSDIS distributes data in granules, which may cover large geographic regions.
- No data services provided.
- Technology insertion continues to improve EOSDIS

Problems in Data-intensive ESSE

- Difficulty to access the huge volume of EOS data.
  - Take weeks to order and obtain a large volume of EOS data.
- Difficulty to use the data.
  - Significant time, resources, and data/IT knowledge are required for preprocessing the multi-source data into a ready-to-analyze form.
  - The ESSE faculty normally does not have enough knowledge in the data/IT knowledge.
- Lack of enough resources to analyze the data.
  - Few universities have the hardware/software resources to handle multi-terabytes of data in simulation and modeling for solving global-scale problems.

Use case: Landslide Model – Risk Assessment and Management

Static Data

- Geology base maps
- Soil type and properties
- Terrain/DEM
- Past earthquake frequencies

Dynamic data

- Land cover map
- Soil moisture (wetness)
- Hydrology
- Precipitation
- Hurricane condition
- Disturbance (construction sites, etc)

Use case: Landslide Model – Risk Assessment and Management (cont.)

Landslide risk modeling:

- Binary method (1 * 1 * 0 * 0…)
- Ranking (1+0+1+0+…)
- Rating method (4+7+3+…)
- Weighted rating (4*2+7*1+3*1+…)
- Other models: R=f(x1,x2,…)

Use case: Landslide Model – Risk Assessment and Management (cont.)

- Stability index map
- Potentially unstable zones
- Informed stability management
- Potential damage
  - Transportation
  - Business/Industrial/etc infrastructures
  - Residential
  - Lakes/Reservoirs/river networks
  - Environmental
  - Ecological/biodiversity
- Potential damage assessment
- Potential damage management

Use case: Landslide Model – Risk Assessment and Management (cont.)

Characteristics of the study:

- Dynamic in nature
- Quick assessment and response essential
- Distributed data sources
- Significantly different data types
- Heterogeneous data formats
- Tremendous data preprocessing
- Model being either simple or complicated
- Chains of data/services involved

Expected Significances

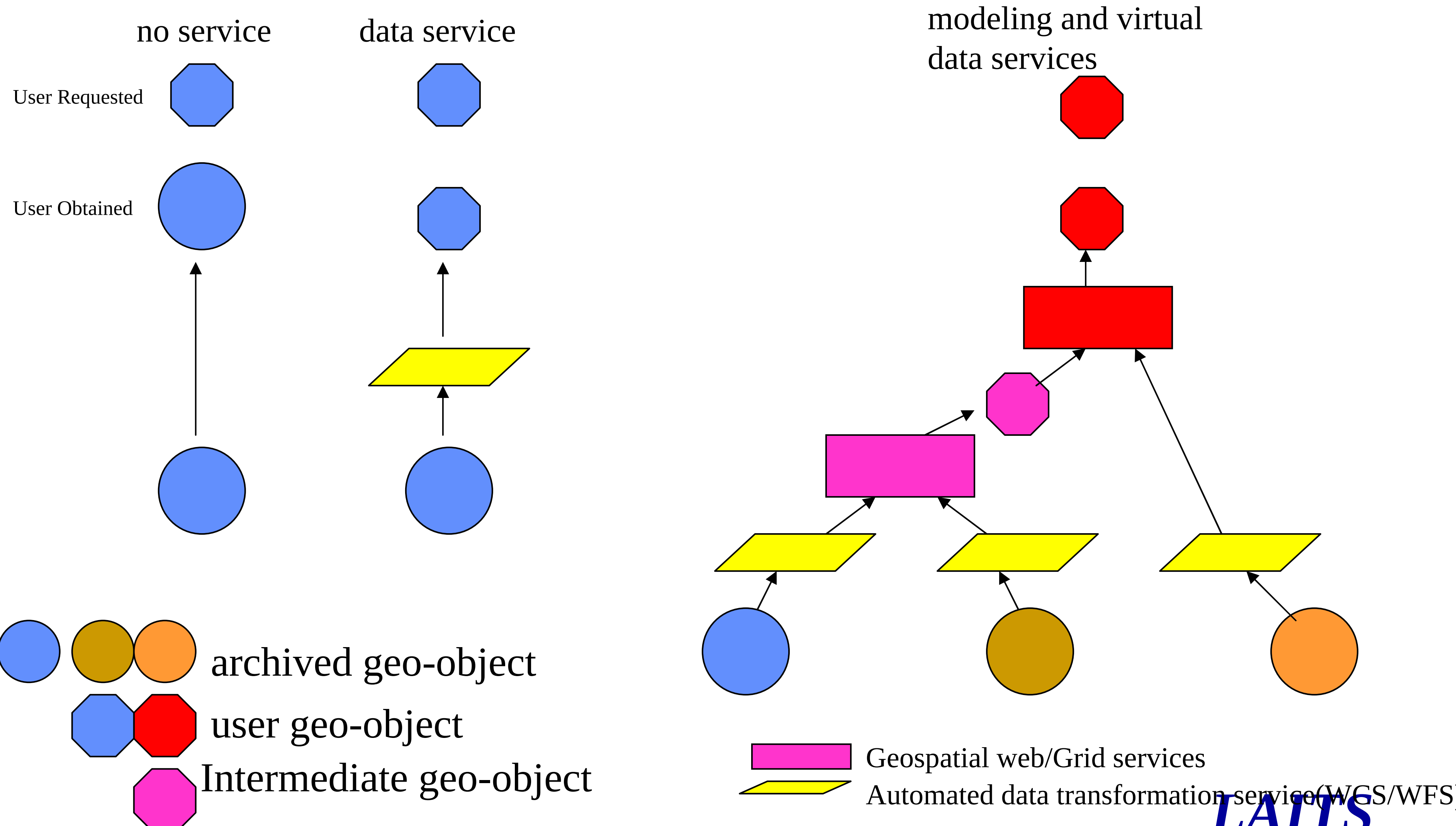
- The GeoBrain system will give ESSE institutes a geospatial data-rich learning and research environment that was never available to them before.
- The environment will enable students interactively, through their desktop computers, explore answers to the scientific questions by mining the peta-bytes of EOSDIS data.
- The technology also provides the interactive collaboration among students worldwide on scientific modeling, knowledge exchanges, and scientific criticism.
- Such an environment will inspire students’ curiosity on sciences and enable faculties and students doing many new studies that could not be done before.
- It will also provide educators with unique teaching tools and compelling teaching experiences that they never have experienced and that only NASA can offer.

Virtual datasets

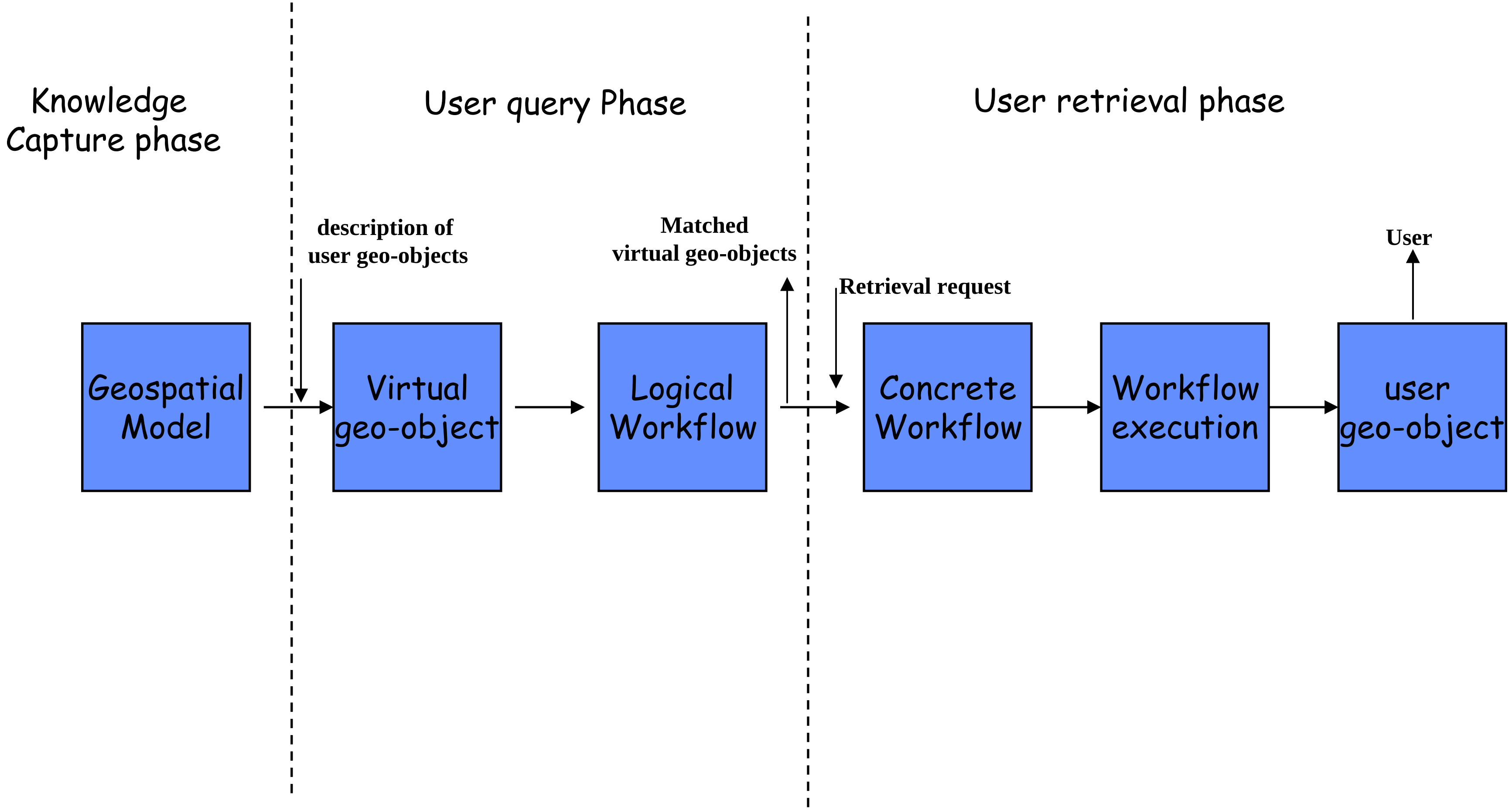
- A virtual dataset is a dataset that:
  - Does not exist in a data and information system
  - The system knows how to create it on-demand.
  - A virtual dataset, once created, can be kept for fulfilling the same request from next users.
- The client/data user will not know the difference between a real dataset and a virtual dataset.
- A virtual dataset can be produced (materialized) by
  - running a computer program dedicated to the production of the virtual dataset (dedicated program approach).
  - running a series of service modules, each one takes care of a small step of the materialization of the virtual dataset (service approach).

George Mason University

Geo-object, Geo-tree, Virtual Dataset, Geospatial Models



From Geospatial model to user geo-object



Knowledge capture: the construction of geospatial models

- Two catalogs are essential
  - geo-object types
  - service types
- Save the model in a formal way
  - Use the modified/simplified BPEL to store the geo-tree
  - Keep geo-trees in the geo-tree library
  - Catalog the geo-trees in the geo-object catalog.
- Two ways to create geospatial models
  - Domain experts create models and share with others.
    - Easy to construct a model through a graphic model construction client.
  - Automatically creation of model.
    - Require the system to have domain knowledge and AI capabilities

User Creation of Geospatial Models

- A user-requested geo-object may not exist either virtually and nor physically.
- If the user knows the process to create the geo-object from lower-level inputs step-by-step (the logical geospatial modeling)
  - With help of a good user interface and the availability of service modules and models/submodels, the user can construct a geospatial model/virtual data product interactively.
  - The system then can produce the virtual data product for the user.
  - The user-created model can be incorporated into the system as a part of the virtual datasets the system can provide.
- This allows the system to grow capabilities with time.
- Advantages
  - allows users to obtain the ready-to-use scientific information instead of the raw data, significantly reducing the data traffic between the users and the geospatial Grid.
  - allows users to explore huge resources available at a data Grid and to conduct tasks that they never be able to conduct before.

Automatic Creation of Geospatial Model

- We are studying the approach to automatically create geospatial process models based on user's description on user-geo-object and produce the user geo-object.
  - In many cases, a model used to generate user requested geo-object does not exist and the user is not an expert user knowing how to create a model.
  - The automatic creation will allow the general users to obtain geospatial information and knowledge
- The steps are as following:
  - Users express their request in nature language (ideally) or in controlled vocabularies.
  - The request is converted to a user geo-object.
  - The user geo-object is abstracted to become a user geo-object type.
  - Deductive method is used to form the geo-tree from the user geo-object type with the help of geospatial ontology.
  - The rests are the same as the other modeling approaches.

From Geo-object to virtual user geo-object

- The root node of a geo-tree is a virtual geo-object type
- When user/client requests a geo-object, user will provide a description of the geo-object they want;
- If the type of the user geo-object matches with virtual geo-object type in a geo-tree,
  - The geo-tree is selected;
  - The root node is instantiated with the descriptions provided by the clients.
  - The root node now becomes a virtual user geo-object.
- The next step is to determine if the virtual user-geo-object can be materialized on the fly.

Logical Instantiation of Geo-Tree

- Check if a virtual user geo-object can be materialized by instantiating the whole geo-tree.
  - Push the description of the virtual user geo-object down to each node of geo-tree (e.g., spatial coverage, format, etc);
  - Discover instance of service and geo-objects through searching both service instance catalog and geo-object instance catalog;
  - If an archive geo-object is found as the input of a process, then the push down will be stop for the branch of this tree.
- The logical instantiation will not create an actual workflow, but conceptually, it creates a logical workflow.
- If a geo-tree can be instantiated logically, the virtual user geo-object can be materialized.
  - A logical ID will be created and return to client to indicate the user-requested geo-object is found in the system.
  - The logical ID will be used by the client/user to request for the geo-object.

Physical Instantiation of Geo-Tree

- When client requests a user geo-object, the geo-object ID will indicate if the user geo-object is virtual.
- If the geo-object is virtual, the geo-tree associated with the virtual geo-object will be instantiated to create a concrete workflow.
  - A workflow language will be used to encode the workflow.
  - The workflow is executable in a workflow execution engine.

Creation of user geo-object

- The workflow engine will execute the workflow and generate the user geo-object.
- The user geo-object will be return to user/client.
- The above mentioned steps reflects two stage processes:
  - User query
  - User retrieve
- The two stages can also be merged into one stage process that when a user query can meet, the resulted user-object will be pushed back to user automatically without user initiation of the retrieval.

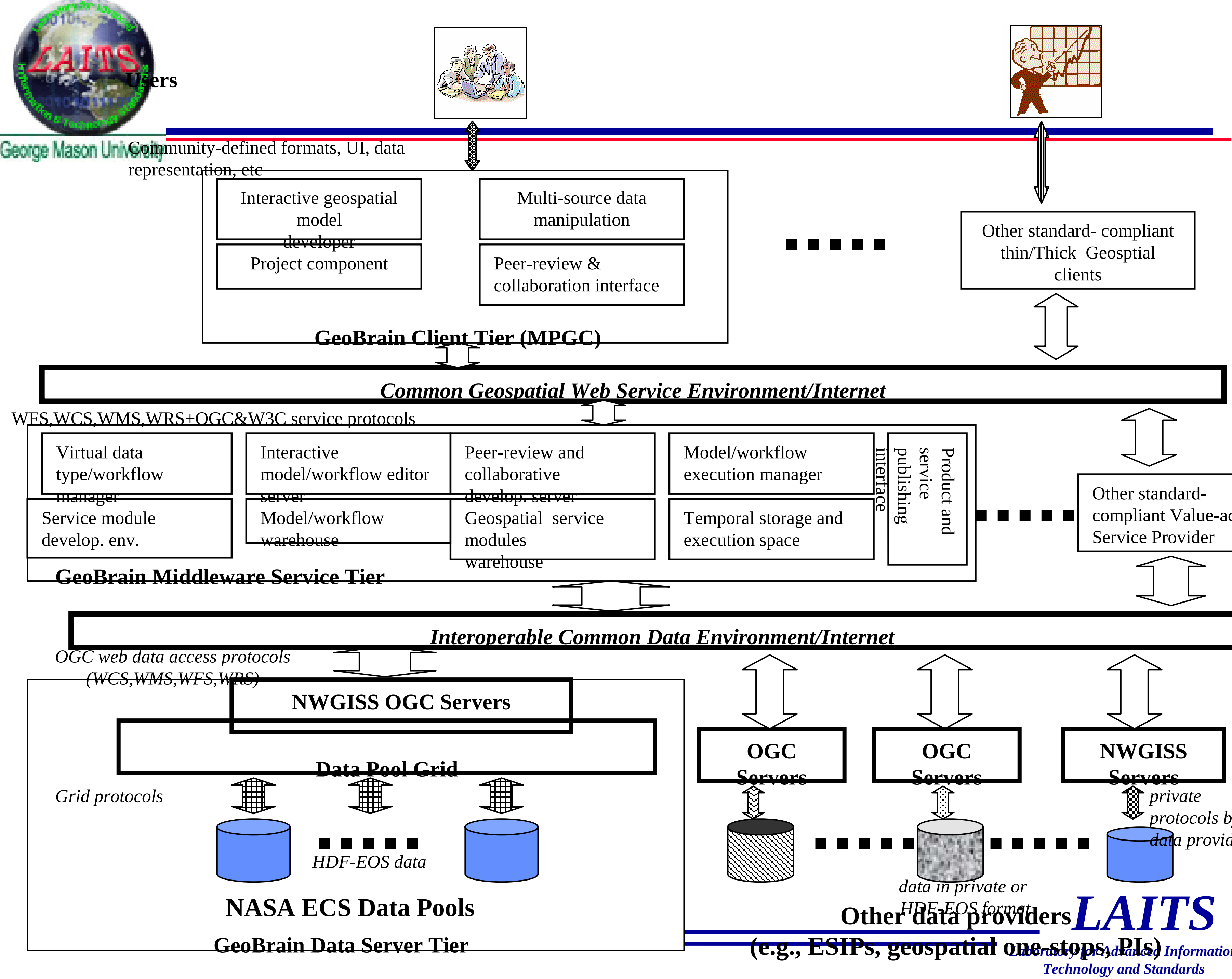
The Infrastructure Foundation

- NASA ESE is working on putting ESS data at DAACs on-line for rapid access through data pools
  - Most commonly requested and most recently acquired data currently.
  - 4 DAACs have data pools online already.
  - Eventually all data will be on-line.
- NASA ESE has excellent network infrastructure for data traffic
  - In most cases, 1Gb/second links between NASA DAACs/research centers.
- NASA ESE has huge computational resources.
- Make the vast data and computational resources available and easily accessible to ESSE institutions

George Mason University

The Technology Foundation

- The web-based geospatial interoperability technology.
  - Standards developed by FGDC, ISO, and OGC.
  - The common interfaces to data archives of different data providers for obtaining personalized ready-to-analyze dataset.
- The web service technology
  - The fundamental technology for E-commence.
  - Web Services are self-contained, self-describing, modular applications that can be published, located, and dynamically invoked across the Web.
  - Automatically and dynamically chaining individual services and connecting services to data for solving complex problems are the goal of semantic web.
- Grid technology
  - Securely share the geographically distributed data and computational resources.


Users
Community-defined formats, UI, data representation, etc
Interactive geospatial model developer
Multi-source data manipulation
Project component
Peer-review & collaboration interface
GeoBrain Client Tier (MPGC)
Other standard- compliant thin/Thick Geosptial clients
Common Geospatial Web Service Environment/Internet
WFS,WCS,WMS,WRS+OGC&W3C service protocols
Virtual data type/workflow manager
Interactive model/workflow editor server
Peer-review and collaborative develop. server
Model/workflow execution manager
Product and service publishing interface
Service module develop. env.
Model/workflow warehouse
Geospatial service modules warehouse
Temporal storage and execution space
GeoBrain Middleware Service Tier
Other standard-compliant Value-added Service Provider
Interoperable Common Data Environment/Internet
OGC web data access protocols (WCS,WMS,WFS,WRS)
NWGISS OGC Servers
Data Pool Grid
Grid protocols
HDF-EOS data
NASA ECS Data Pools
GeoBrain Data Server Tier
OGC Servers
OGC Servers
NWGISS Servers
private protocols by data providers
data in private or HDF-EOS format
Other data providers
(e.g., ESIPs, geospatial one-stops, PIs)
George Mason University
LAITS
Laboratory for Advanced Information Technology and Standards

System requirement at the user-side

- Any internet connected PC capable of running JAVA client of the system.
  – The client will be provided to any users for free.
- No fast network connection is required
  – all data reduction is done by the system at computers that users don't need to know.
  – Users only get the result back instead of all raw data.
- No powerful computer with large disk storage capability is needed
  – Basically the users possess the huge computational and data resources that the system can mobilize.
- No expensive analysis software is needed
  – Analysis and modeling capabilities are provided by the system

System built by ESSE community for the community

- The GeoBrain system will be built by the ESS higher-education community for the community.
- The major tasks of system development will be:
  - Development of service framework that allows the automated execution of services and service chains.
  - Development of services modules and geospatial models.
- Any individuals can contribute both modules and models.
- A peer-review panel will be set up to review and validate the modules and models contributed by the community.

Involvement of ESSE Community

- As the users of the system.
  - Provide the requirements
  - Evaluate the systems
  - Develop new curriculums and research around the newly available capabilities.
- Participate in the system development
  - Develop individual service modules
  - Contribute the geospatial modules

Evolution and Self-enhancement of the System

- Beside the computational and network capacity and the data holdings in various distributed archives, the power of the system relies on the availability of the service modules and geospatial models.
- With more and more contributions of modules and models from the user community, the system will become more and more powerful and knowledgeable.
- The inclusions of the modules and models into the system will be subjected to rigorous peer review and testing.

George Mason University

Sharing Technology with other REASoN Teams

- Technology available to other teams
  - OGC interoperable data access technology
    - WCS server
    - WMS server
    - WRS/Catalog server
    - Multiple-protocol Geoinformation Client
    - HDF-EOS/GIS translators
  - Technical support on geospatial standards and specifications
- Joint technology development
  - Dynamic model composition through decomposition (implementation of the geotree concept)
  - Workflow management and executions
  - Interoperability of geospatial processes
  - Geospatial web service technology
- Availability of OGC compliant data access and services
  - Serve EOS data using OGC protocols.
  - Can be used in testbeds to test interoperability.

The Team

- Development Team
  - George Mason University
  - City University of New York
  - Northern Illinois University
  - University of Texas – Dallas
- Education partners
  - In the first three years of the project, three education partners will be selected in each year through a RFP process (Total 9 partners).
  - Each partner will be provided two years of funds to develop new/enhanced courses based the capabilities, promote the use of the system in the peers, and provide feedback to the development team.
  - Any higher-education professors and students are welcomed to use the system and participate in the development.